Group assignment 2

Assignment_2_21

November 2025

Group Members

Baraa Magdy

Celia Medina Giménez

Dino Keylas

Maria Naz

Oskar Johannes Piibar

Problem 1.

We assume the conditional distribution

$$Y \mid X = x \sim \text{Poisson}(\lambda(x)), \qquad \lambda(x) = \exp(\alpha \cdot x + \beta),$$

where $\alpha \in \mathbb{R}^p$ is a slope vector and $\beta \in \mathbb{R}$ is an intercept. The conditional pmf for a single observation (x, y) is

$$f_{Y|X}(y \mid x) = \frac{\lambda(x)^y e^{-\lambda(x)}}{y!}.$$

Following the Maximum Likelihood viewpoint (Section 4.2.1: Maximum Likelihood and regression), derive the loss to be minimized w.r.t. α, β. Address whether the factorial term is needed.

Log-likelihood and negative log-likelihood

For one observation (x, y) the log-likelihood is

$$\ell(\alpha, \beta \mid x, y) = \log f_{Y|X}(y \mid x) = y \log \lambda(x) - \lambda(x) - \log(y!).$$

Using $\log \lambda(x) = \alpha \cdot x + \beta$ and $\lambda(x) = \exp(\alpha \cdot x + \beta)$ we get

$$\ell(\alpha, \beta \mid x, y) = y(\alpha \cdot x + \beta) - \exp(\alpha \cdot x + \beta) - \log(y!).$$

The negative log-likelihood (loss) for a single observation is therefore

$$\ell_{\text{neg}}(\alpha, \beta \mid x, y) = -y(\alpha \cdot x + \beta) + \exp(\alpha \cdot x + \beta) + \log(y!).$$

For a dataset $\{(x_i, y_i)\}_{i=1}^n$ the total negative log-likelihood is

$$L(\alpha, \beta) = \sum_{i=1}^{n} \big[- y_i(\alpha \cdot x_i + \beta) + \exp(\alpha \cdot x_i + \beta) + \log(y_i!)\big].$$

Practical loss to minimize

The term $\sum_{i=1}^n \log(y_i!)$ does *not* depend on the parameters (α, β). Hence it is a constant for optimization and can be dropped. The practical loss (negative log-likelihood up to an additive constant) to minimize is

$$\boxed{\widetilde{L}(\alpha, \beta) = \sum_{i=1}^{n} \big[\exp(\alpha \cdot x_i + \beta) - y_i(\alpha \cdot x_i + \beta)\big].}$$

Thus *we do not need the factorial term* when performing parameter estimation by MLE, because it does not affect the location of the minimizer.

Score (first-order) equations

Define $\lambda_i = \exp(\alpha \cdot x_i + \beta)$. The gradient (score) components are

$$\frac{\partial \widetilde{L}}{\partial \alpha} = \sum_{i=1}^{n} (\lambda_i - y_i)\, x_i, \qquad \frac{\partial \widetilde{L}}{\partial \beta} = \sum_{i=1}^{n} (\lambda_i - y_i).$$

Setting these to zero yields the likelihood equations

$$\sum_{i=1}^{n} (\lambda_i - y_i)\, x_i = 0, \qquad \sum_{i=1}^{n} (\lambda_i - y_i) = 0,$$

which are the analogues of the normal equations for Poisson regression.

Hessian and convexity

Second derivatives give the Hessian blocks

$$\frac{\partial^2 \widetilde{L}}{\partial \alpha \partial \alpha^\top} = \sum_{i=1}^{n} \lambda_i\, x_i x_i^\top, \quad \frac{\partial^2 \widetilde{L}}{\partial \beta^2} = \sum_{i=1}^{n} \lambda_i, \quad \frac{\partial^2 \widetilde{L}}{\partial \alpha \partial \beta} = \sum_{i=1}^{n} \lambda_i\, x_i.$$

Since $\lambda_i > 0$ for all i, the Hessian is positive semidefinite (in typical cases positive definite), so $\widetilde{L}$ is convex in (α, β) and standard convex optimization methods apply.

Connection to GLM / IRLS

Minimizing $\widetilde{L}$ is exactly maximum likelihood estimation for a Generalized Linear Model with Poisson family and log link. Typical solvers use Newton / Fisher scoring / IRLS (iteratively reweighted least squares), where in each iteration weights proportional to λ_i appear.

Summary (final answer)

The loss to minimize (negative log-likelihood up to constant) is

$$\boxed{\widetilde{L}(\alpha, \beta) = \sum_{i=1}^{n} \Big[\exp(\alpha \cdot x_i + \beta) - y_i(\alpha \cdot x_i + \beta) \Big].}$$

The factorial term $\log(y_i!)$ can be omitted during optimization because it does not depend on (α, β).

Problem 2.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from Uniform$(0, \theta)$.

Let $\hat{\theta} = \max(X_1, X_2, \ldots, X_n)$.

(a) Find the distribution function of $\hat{\theta}$.

(b) Compute the bias Bias$(\hat{\theta})$, the standard error SE$(\hat{\theta})$, and the mean squared error MSE$(\hat{\theta})$.

Problem 2a.

CDF of $\hat{\theta}$

We want CDF of

$$\hat{\theta} = \max(X_1, X_2, ..., X_n)$$

That means we need

$$F_{\hat{\theta}}(x) = (\hat{\theta} \leq x)$$

If the maximum $(\hat{\theta})$ is $\leq x$, that means

$$\max(X_1, X_2, ..., X_n) \leq x$$

Hence,

$$P(\hat{\theta} \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x)$$

Because X_i's are independent, then

$$P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = P(X_1 \leq x) \cdot P(X_2 \leq x) \cdot ...P(X_n \leq x)$$

Since each has the same CDF, $F_X(x)$, this becomes

$$P(\hat{\theta} \leq x) = [F_X(x)]^n$$

For a Uniform(0, θ) random variables:

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \left(\frac{x}{\theta}\right), & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

As We know that $P(\hat{\theta} \leq x) = [F_X(x)]^n$, then the **CDF** of $\hat{\theta}$:

$$F_{\hat{\theta}}(x) = \begin{cases} 0, & x < 0, \\ \left(\frac{x}{\theta}\right)^n, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

By differentiating the **CDF**, we can get the **PDF**:

$$\begin{aligned} f_{\hat{\theta}}(x) &= \frac{d}{dx} F_{\hat{\theta}}(x) \\ &= \frac{d}{dx} \left(\frac{x}{\theta}\right)^n \\ &= \frac{n}{\theta^n} x^{n-1}, \quad 0 \leq x \leq \theta \end{aligned}$$

Problem 2b.

Bias:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Calculate $E(\hat{\theta})$:

$$E(\hat{\theta}) = \int_0^\theta x\, f(x)\, dx = \int_0^\theta x \cdot \frac{n}{\theta^n} x^{n-1}\, dx = \frac{n}{\theta^n} \int_0^\theta x^n\, dx$$

$$= \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\,\theta.$$

Plug $E(\hat{\theta})$ to the bias formula:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\,\theta.$$

Standard Error:

$$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

Calculate $\text{Var}(\hat{\theta})$:

$$E(\hat{\theta}^2) = \int_0^\theta x^2 f(x)\, dx = \frac{n}{\theta^n} \int_0^\theta x^{n+1}\, dx = \frac{n}{\theta^n} \cdot \frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\,\theta^2.$$

$$\text{Var}(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2$$

$$= \theta^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) = \theta^2 \cdot \frac{n}{(n+2)(n+1)^2}.$$

Plug $\text{Var}(\hat{\theta})$ to the SE formula:

$$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})} = \theta \sqrt{\frac{n}{(n+2)(n+1)^2}} = \frac{\theta}{n+1}\sqrt{\frac{n}{n+2}}.$$

Mean Squared Error:

$$\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

$$= \left(-\frac{\theta}{n+1}\right)^2 + \frac{n\,\theta^2}{(n+1)^2(n+2)}$$

$$= \frac{\theta^2}{(n+1)^2} + \frac{n\,\theta^2}{(n+1)^2(n+2)}$$

$$= \frac{\theta^2}{(n+1)^2}\left(1 + \frac{n}{n+2}\right)$$

$$= \frac{\theta^2}{(n+1)^2} \cdot \frac{n+2+n}{n+2}$$

$$= \frac{(n+3)\,\theta^2}{(n+1)^2(n+2)}.$$

Problem 3.

Consider the continuous distribution with density

$$p(x) = \frac{1}{2}\cos x, \qquad -\frac{\pi}{2} < x < \frac{\pi}{2},$$

Problem 3a.

Find the distribution function $F(x)$

Solution

By definition,

$$F(x) = \int_{-\infty}^{x} f(v)\, dv.$$

When $x \leq -\frac{\pi}{2} \Rightarrow F(x) = 0$, and when $x \geq \frac{\pi}{2} \Rightarrow F(x) = 1$.

Otherwise,

$$F(x) = \int_{-\pi/2}^{x} \frac{1}{2}\cos(t)\, dt = \frac{1}{2}[\sin(t)]_{-\pi/2}^{x} = \frac{1}{2}(\sin(x) + 1).$$

So,

$$F(x) = \begin{cases} 0, & x \leq -\frac{\pi}{2}, \\ \frac{1}{2}(\sin(x) + 1), & -\frac{\pi}{2} < x < \frac{\pi}{2}, \\ 1, & x \geq \frac{\pi}{2}. \end{cases}$$

Problem 3b.

Find the inverse distribution function $F^{-1}(u)$

Solution

From part (a),

$$F(x) = \frac{1}{2}(\sin(x) + 1).$$

So we set

$$u = F(x) = \frac{1}{2}(\sin(x) + 1)$$

Now, let's just solve for x.

$$2u - 1 = \sin(x) \Rightarrow x = \arcsin(2u - 1).$$

Therefore,

$$F^{-1}(u) = \arcsin(2u - 1), \quad 0 < u < 1.$$

Problem 3c.

To sample using an *Accept–Reject sampler* (Algorithm 1), we need to find a density $g(x)$ such that

$$p(x) \leq Mg(x) \quad \text{for some } M > 0.$$

Find such a density g and determine the corresponding value of M.

Solution

p(x) is bounded and has an "easy" shape (symmetric, has a maximum at x=0, and goes to 0 at its boundaries). This means an uniform distribution on the same interval seems appropriate. So let g(x):

$$g(x) = \frac{1}{b-a} = \frac{1}{\pi}, \qquad x \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big].$$

Now, let's check that it is a valid pdf:

$$\int_a^b g(x)\,dx = \int_{-\pi/2}^{\pi/2} \frac{1}{\pi}\,dx = \frac{1}{\pi}(\pi) = 1.$$

Hence, $g(x)$ is a valid pdf.
We require that $p(x) \leq Mg(x)$ for all x.

$$\frac{p(x)}{g(x)} \leq M \Rightarrow \frac{p(x)}{g(x)} = \frac{\frac{1}{2}\cos(x)}{\frac{1}{\pi}} = \frac{\pi}{2}\cos(x)$$

We look for the maximum of $\frac{p(x)}{g(x)}$ at $x = 0$ to find the lowest value of M that would "cover" all p(x) with g(x)M.

$$M = \frac{\pi}{2}\cos(0) = \frac{\pi}{2}.$$

Thus:

$$M = \frac{\pi}{2}, \quad g(x) = \frac{1}{\pi}.$$

This makes sense graphically, since $Mg(x)$ just covers $p(x)$ at its maximum.

Problem 4.

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of IID discrete random variables, where $P(Y_i = 0) = 0.1$, $P(Y_i = 1) = 0.3$, $P(Y_i = 2) = 0.2$, and $P(Y_i = 3) = 0.4$. Let $X_n = \max\{Y_1, Y_2, \ldots, Y_n\}$. Let $X_0 = 0$ and verify that $(X_0, X_1, \ldots, X_n)$ is a Markov chain. Find the transition matrix P.

Solution

To show that $(X_0, X_1, \ldots, X_n)$ is a Markov chain, we have to prove that it has at most 1-step memory.

For a step X_n in the chain, we can say that:

$$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0).$$

From the problem definition, we know that X_n is the maximum of all the previous Y's:

$$X_n = \max\{Y_1, Y_2, \ldots, Y_n\},$$

which can be rewritten as

$$\max\{X_{n-1}, Y_n\},$$

because

$$(X_1, X_2, \ldots, X_{n-1}) = (X_1 = \max\{Y_1\}, X_2 = \max\{Y_1, Y_2\}, \ldots, X_{n-2} = \max\{Y_1, Y_2, \ldots, Y_{n-2}\}, X_{n-1} = \max\{Y_1, Y_2, \ldots, Y_{n-1}\}).$$

Now, rewriting the initial statement gives us:

$$\mathbb{P}(\max\{X_{n-1}, Y_n\} = x_n \mid \max\{X_{n-2}, Y_{n-1}\} = x_{n-1}, \max\{X_{n-3}, Y_{n-2}\} = x_{n-2}, \ldots, \max\{X_0, Y_1\} = x_1, X_0 = x_0).$$

However, since the sequence of Y's is independent, this means that Y_n is independent of previous $(Y_1, Y_2, \ldots, Y_{n-1})$ and also independent of $(X_1, X_2, \ldots, X_{n-1})$. Therefore, conditioning on earlier X's beyond X_{n-1} does not affect the distribution of X_n. Taking this into account, we can write the last statement as:

$$\begin{aligned}\mathbb{P}(\max\{X_{n-1}, Y_n\} = x_n \mid \max\{X_{n-2}, Y_{n-1}\} = x_{n-1}, \max\{X_{n-3}, Y_{n-2}\} = x_{n-2}, \ldots, \max\{X_0, Y_1\} = x_1, X_0 = x_0) \\ = \mathbb{P}(\max\{X_{n-1}, Y_n\} = x_n \mid \max\{X_{n-2}, Y_{n-1}\} = x_{n-1}) \\ = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}),\end{aligned}$$

which is the definition of being a Markov chain. This concludes verifying that $(X_0, X_1, \ldots, X_n)$ is a Markov chain.

To construct the transition matrix P we begin by identifying the dimensions of the matrix. As Y can take 4 different values - {0, 1, 2, 3 } - the matrix will be a 4x4 matrix. This is because we have 4 states to begin from (the rows) and 4 states to go to (the columns).

Now to fill it with values the following logic was applied: for each entry P_{ij} in matrix P, where i indicates the row and j the column, both in the range (0-3), conditional probability was applied. For example, the row $P_{00}, P_{01}, P_{02}, P_{03}$ was filled out with the probabilities given in the exercise, which indicates going from state $X_{n-1} = 0$ to $X_n = 0, 1, 2, 3$, respectively. Let us assume that we are now at state $X_n = 1$ and as $X_{n+1} = \max\{X_n, Y_{n+1}\}$ then returning to state 0 is no longer possible ($P_{10} = 0$). Now P_{11} equals the probability that $Y_i = 0$ or $Y_i = 1 \Rightarrow 0.1 + 0.3 = 0.4$. P_{12} and P_{13} are filled in a similar way to the fields P_{02}, P_{03}. For the row vector $P_{20}, P_{21}, P_{22}, P_{23}$, we can no longer reach the states 0, 1, therefore $P_{20}, P_{21} = 0$ and $P_{22} = \mathbb{P}(Y_i = 0) + \mathbb{P}(Y_i = 1) + \mathbb{P}(Y_i = 2) = 0.1 + 0.3 + 0.2 = 0.6$. Reaching state 3 from initial state 2 has a probability of 0.4, given in the exercise description. For the final row vector $P_{30}, P_{31}, P_{32}, P_{33}$, we can no longer reach any of the states other than 3, therefore $P_{30}, P_{31}, P_{32} = 0$ and $P_{33} = 0.1 + 0.3 + 0.2 + 0.4 = 1$. The completed transition matrix P can be seen below:

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0 & 0.4 & 0.2 & 0.4 \\ 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The rows of a transition matrix have to sum up to 1, which is the case for the constructed matrix, giving reasons to believe that it was filled properly.

Problem 5.

Let $X_1, \ldots, X_n$ be IID from an unknown distribution F. Let $\widehat{F}_n$ be the empirical distribution function. Use this to find an estimate of the p-quantile of F (call it q). Use Theorem 5.28 to find a confidence interval for q.

Solution.

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with empirical CDF

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}, \qquad X_{(1)} \le \cdots \le X_{(n)} \text{ (order statistics).}$$

For a fixed $p \in (0,1)$ the population p-quantile (Remark 6.15) is:

$$q_p := F^{-1}(p) = \inf\{x : F(x) \ge p\}.$$

The empirical p-quantile is the inverse of $\widehat{F}_n$:

$$\widehat{q}_p := \widehat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}.$$

Here np is the product of the sample size n and the probability level p; it is the expected rank (i.e., the expected number of observations $\le q_p$) in a sample of size n. Taking $\lceil np \rceil$ chooses the smallest index whose cumulative proportion k/n is at least p.

Confidence interval via DKW (Theorem 5.28).

The Dvoretzky–Kiefer–Wolfowitz inequality says that for any $\varepsilon > 0$,

$$\Pr\Big(\sup_x \big|\widehat{F}_n(x) - F(x)\big| > \varepsilon \Big) \le 2e^{-2n\varepsilon^2}.$$

For a confidence level $1-\alpha$, set

$$\varepsilon_\alpha := \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

Then, with probability at least $1-\alpha$,

$$q_p \in \big[\widehat{F}_n^{-1}(p - \varepsilon_\alpha),\ \widehat{F}_n^{-1}(p + \varepsilon_\alpha)\big].$$

Since $\widehat{F}_n^{-1}(u) = X_{(\lceil nu \rceil)}$, this becomes the explicit interval

$$\mathrm{CI}_{1-\alpha}(q_p) = \Big[X_{(\,\lceil n\,\underline{p}\rceil\,)}\ ,\ X_{(\,\lfloor n\,\overline{p}\rfloor\,)}\Big], \qquad \underline{p} = \max\{0, p-\varepsilon_\alpha\}, \quad \overline{p} = \min\{1, p+\varepsilon_\alpha\}.$$